利用神经网络进行序列到序列学习

Ilya Sutskever, Oriol Vinyals, Quoc V. Le

摘要：深度神经网络（DNNs）是强大的模型，已在困难的学习任务上取得了卓越的表现。尽管 DNNs 在大型标记训练集可用时表现良好，但它们不能用于序列到序列的映射。在本文中，我们提出了一种对序列学习的通用端到端方法，这种方法对序列结构的假设最小。我们的方法使用多层长短期记忆（LSTM）将输入序列映射到一个固定维度的向量，然后再使用另一个深层 LSTM 从该向量解码目标序列。我们的主要结果是，在 WMT-14 数据集上的一个英语到法语的翻译任务上，LSTM 生成的翻译在整个测试集上达到了 34.8 的 BLEU 分数，其中 LSTM 的 BLEU 分数的惩罚规则为出现词汇表外的词。此外，LSTM 在处理长句子时没有困难。作为比较，一个基于短语的 SMT 系统在同一数据集上达到了 33.3 的 BLEU 分数。当我们使用 LSTM 重新排列前述 SMT 系统产生的 1000 个假设时，其 BLEU 分数提高到了 36.5，这接近于先前的最佳水平。LSTM 还学会了敏感于词序并且相对不变于主动和被动语态的合理短语和句子表示。最后，我们发现，反转所有源句子（但不是目标句子）中的词序显著提高了 LSTM 的性能，因为这样做在源句子和目标句子之间引入了许多短期依赖，这使得优化问题更容易。

关键词：DNNs，LSTM，BLUE 分数、文本翻译

1 引言

深度神经网络（DNNs）是极其强大的机器学习模型，在诸如语音识别[13, 7]和视觉对象识别[19, 6, 21, 20]等困难问题上取得了出色的表现。DNNs 之所以强大，是因为它们可以在适度数量的步骤中执行任意的并行计算。DNNs 强大的一个令人惊讶的例子是，它们仅使用两个隐藏层的二次大小就能对 N 个 N 位数进行排序[27]。因此，虽然神经网络与传统的统计模型有关，它们学习了复杂的计算。此外，只要标记训练集有足够的信息来指定网络的参数，就可以使用有监督的反向传播来训练大型 DNNs。因此，如果存在一种大型 DNN 的参数设置可以取得良好结果（例如，因为人类可以非常迅速地解决任务），有监督的反向传播将找到这些参数并解决问题。

尽管 DNNs 灵活且强大，但它们只能应用于输入和目标可以用固定维度的向量合理编码的问题。这是一个重大限制，因为许多重要的问题最好用长度未知的序列来表达。例如，语音识别和机器翻译是序列问题。同样，问答也可以看作是将表示问题的一系列单词映射到表示答案的一系列单词。因此，显然，一种学习将序列映射到序列的领域独立方法将是有用的。

序列对 DNNs 构成挑战，因为它们要求输入和输出的维度是已知且固定的。在本文中，我们展示了直接应用长短期记忆（LSTM）架构[16]可以解决一般的序列到序列问题。

思路是使用一个 LSTM 逐个时间步读取输入序列，以获得大的固定维度向量表示，然后使用另一个 LSTM 从该向量中提取输出序列（图 1）。第二个 LSTM 本质上是一个循环神经网络语言模型[28, 23, 30]，只不过它是以输入序列为条件的。LSTM 成功学习具有长范围时间依赖性的数据的能力，使其成为这一应用的自然选择，这是由于输入及其相应输出之间的相当大的时间延迟（图 1）。

已经有许多相关尝试通过神经网络解决一般序列到序列学习问题。我们的方法与 Kalchbrenner 和 Blunsom[18]密切相关，他们是首次将整个输入句子映射到向量，并且与 Cho 等人[5]非常相似。Graves[10]引入了一种新颖的可微注意力机制，允许神经网络关注其输入的不同部分，Bahdanau 等人[2]成功地将这一理念的一个优雅变体应用于机器翻译。连接主义序列分类是另一种流行的技术，用于通过神经网络将序列映射到序列，尽管它假设输入和输出之间有一个单调对齐[11]。

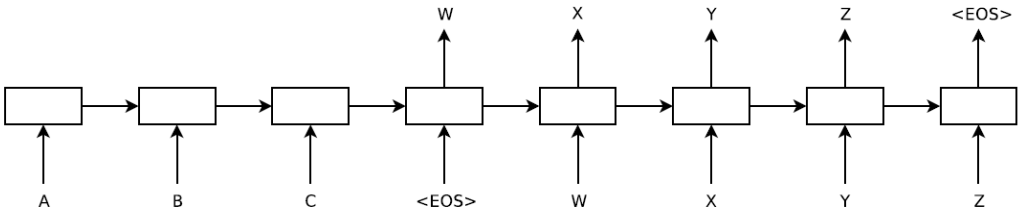

图 1 我们的模型读取输入句子“ABC”，并产生“WXYZ”作为输出句子。在输出句子结束标记后，模型停止做出预测。值得注意的是，LSTM 以相反的顺序读取输入句子，因为这样做在数据中引入了许多短期依赖性，使得优化问题变得更加容易。

本工作的主要结果如下。在 WMT'14 英语到法语翻译任务上，我们通过使用一个简单的从左到右的 beam-search 解码器，直接从 5 个深层 LSTMs（每个有 3.8 亿参数）的集成中提取翻译，获得了 34.81 的 BLEU 分数。这是迄今为止通过大型神经网络直接翻译所获得的最好结果。作为比较，该数据集上一个 SMT 基线的 BLEU 分数是 33.30[29]。34.81 的 BLEU 分数是由一个词汇量为 80k 单词的 LSTM 达成的，因此每当参考翻译包含这 80k 个词汇表外的词时，分数会受到惩罚。这个结果表明，一个相对未优化的神经网络架构，仍有很大的改进空间，却已经超越了成熟的基于短语的 SMT 系统。

最后，我们使用 LSTM 对同一任务上 SMT 基线的公开可用的 1000 最佳列表进行了重新打分[29]。通过这样做，我们获得了 36.5 的 BLEU 分数，这提高了基线 3.2 个 BLEU 点，接近先前的最佳水平（即 37.0[9]）。

令人惊讶的是，尽管其他研究人员在相关架构上的最近经验中遇到了问题，LSTM 在处理非常长的句子时并没有遇到困难。我们能够很好地处理长句子，是因为我们在训练和测试集中反转了源句子中的单词顺序，但没有反转目标句子中的单词顺序。通过这样做，我们引入了许多短期依赖，这使得优化问题大为简化（见第 2 节和 3.3 节）。因此，SGD 能够学习没有长句子问题的 LSTMs。在源句子中反转单词的简单技巧是本工作的关键技术贡献之一。

LSTM 的一个有用属性是，它学会了将可变长度的输入句子映射到固定维度的向量

表示中。鉴于翻译倾向于是源句子的释义，翻译目标鼓励 LSTM 找到捕捉它们含义的句子表示，因为具有相似含义的句子彼此接近，而不同含义的句子则会远离。定性评估支持这一说法，显示我们的模型意识到词序，并且对主动和被动语态相当不变。

2 模型

递归神经网络（RNN）[31, 28]是前馈神经网络向序列的自然泛化。给定一系列输入（$x_1, \dots, x_T$），一个标准的 RNN 通过迭代以下方程计算一系列输出（$y_1, \dots, y_T$）：

$$h_t = \sigma(W^{hx} x_t + W^{hh} h_{t-1})$$

$$y_t = W^{yh} h_t$$

当输入与输出之间的对齐提前已知时，RNN 可以轻松地将序列映射到序列。然而，当输入和输出序列的长度不同，且关系复杂且非单调时，如何应用 RNN 到这类问题上还不清楚。

一种通用序列学习的简单策略是使用一个 RNN 将输入序列映射到一个固定大小的向量，然后用另一个 RNN 将该向量映射到目标序列（Cho 等人[5]也采用了这种方法）。原则上这是可行的，因为 RNN 被提供了所有相关信息，但由于长期依赖性的结果[14, 4]（图 1）[16, 15]，训练 RNN 将会很困难。然而，长短期记忆（LSTM）[16]已知能学习具有长期时间依赖性的问题，因此 LSTM 在这种设置下可能会成功。

LSTM 的目标是估计条件概率$p(y_1, \dots, y_{T'} | x_1, \dots, x_T)$，其中$(x_1, \dots, x_T)$是输入序列，$y_1, \dots, y_{T'}$是其对应的输出序列，其长度$T'$可能与$T$不同。LSTM 通过首先获得输入序列$(x_1, \dots, x_T)$的最后一个隐藏状态给出的固定维度表示$v$，然后使用标准的 LSTM-LM 公式计算$y_1, \dots, y_{T'}$的概率，其初始隐藏状态被设置为$x_1, \dots, x_T$的表示$v$：

$$p(y_1, \dots, y_{T'} | x_1, \dots, x_T) = \prod_{t=1}^{T'} p(y_t | y_1, \dots, y_{t-1}) \quad (1)$$

在这个方程中，每个$p(y_t | v, y_1, \dots, y_{t-1})$分布用词汇表中所有单词上的 softmax 表示。我们使用了 Graves[10]的 LSTM 公式。注意，我们要求每个句子以一个特殊的句子结束符号“<EOS>”结束，这使得模型能够定义所有可能长度序列的分布。整个方案在图 1 中概述，所展示的 LSTM 计算了“A”、“B”、“C”、“<EOS>”的表示，然后使用这个表示来计算“W”、“X”、“Y”、“Z”、“<EOS>”的概率。

我们的实际模型在三个重要方面与上述描述不同。首先，我们使用了两个不同的 LSTM：一个用于输入序列，另一个用于输出序列，因为这样做在可忽略的计算成本下增加了模型参数的数量，并使得自然地同时对多种语言对进行 LSTM 训练成为可能[18]。其次，我们发现深层 LSTM 在性能上显著超过浅层 LSTM，因此我们选择了一个四层的 LSTM。第三，我们发现反转输入句子中单词的顺序极其有价值。例如，不是将句子 a, b, c 映射到句子α, β, γ，而是让 LSTM 将 c, b, a 映射到α, β, γ，其中α, β, γ是 a, b, c

的翻译。这样做的结果是 a 与α的距离很近，b 与β相当接近等等，这一事实使得 SGD 很容易“建立起”输入和输出之间的“通信”。我们发现这种简单的数据转换大大提升了 LSTM 的性能。

3 实验

我们以两种方式将我们的方法应用到 WMT'14 英语到法语的机器翻译（MT）任务中。我们使用它来直接翻译输入句子，而不使用参考 SMT 系统，并且用它对 SMT 基线的 n-best 列表进行重打分。我们报告了这些翻译方法的准确性，展示了样本翻译，并可视化了得到的句子表示。

3.1 数据集细节

我们使用了 WMT'14 英语到法语数据集。我们在由 348M 法语词和 304M 英语词组成的 1200 万句子的子集上训练我们的模型，这是来自[29]的一个干净的“精选”子集。我们选择这个翻译任务和这个特定的训练集子集，是因为它们的标记化训练和测试集与 SMT 基线的 1000-best 列表都是公开可用的[29]。

由于典型的神经语言模型依赖于每个词的向量表示，我们对两种语言都使用了固定的词汇表。我们对源语言使用了最常见的 160,000 个单词，对目标语言使用了最常见的 80,000 个单词。每个超出词汇表的单词都被替换成一个特殊的“UNK”标记。

3.2 解码和重打分

我们实验的核心是训练一个大型的深层 LSTM 在众多的句子对上。我们通过最大化给定源句子 S 时正确翻译 T 的对数概率来训练它，所以训练目标是

$$\frac{1}{|S|}\sum_{(T,S)\in S}\log p\,(T|S)$$

其中 S 是训练集。一旦训练完成，我们通过寻找根据 LSTM 最可能的翻译来产生翻译：

$$\hat{T} = \arg\max_{T} p\,(T|S) \tag{2}$$

我们使用一个简单的从左到右的 beam search 解码器来寻找最可能的翻译，该解码器维护了一个小数量 B 的部分假设，其中部分假设是某个翻译的前缀。在每个时间步骤，我们将 beam 中的每个部分假设用词汇表中的每个可能的词来扩展。这大大增加了假设的数量，因此我们根据模型的对数概率丢弃除了 B 个最可能的假设之外的所有假设。一旦“<EOS>”符号被附加到一个假设上，它就会从 beam 中移除，并且被添加到完整假设的集合中。虽然这个解码器是近似的，但它简单易实施。有趣的是，即使在 beam 大小为 1 的情况下，我们的系统也能表现良好，且大小为 2 的 beam 提供了 beam search

的大部分好处（表 1）。

我们还使用 LSTM 重新打分基线系统[29]产生的 1000 个最佳列表。为了重新打分一个 n 最佳列表，我们计算了我们的 LSTM 对每个假设的对数概率，并将其分数与基线系统的分数取了平均。

3.3 反转源句子

虽然 LSTM 有能力解决具有长期依赖的问题，我们发现当源句子被反转时（目标句子不反转），LSTM 学得更好。通过这样做，LSTM 的测试困惑度从 5.8 降到了 4.7，其解码翻译的测试 BLEU 分数从 25.9 增加到了 30.6。

虽然我们没有对这一现象的完整解释，我们认为这是由于在数据集中引入了许多短期依赖所致。通常情况下，当我们将源句子与目标句子连接起来时，源句子中的每个单词与目标句子中的对应单词距离都很远。结果是，这个问题有一个较大的“最小时间滞后”[17]。通过反转源句子中的单词，源语言和目标语言中对应单词之间的平均距离没有变化。然而，源语言中的前几个单词现在与目标语言中的前几个单词非常接近，所以问题的最小时间滞后大大减少。因此，反向传播在“建立”源句子和目标句子之间的“通信”上更加轻松，这反过来又显著提高了整体性能。

起初，我们认为反转输入句子只会导致目标句子前部分的预测更加自信，而后部分的预测不那么自信。然而，在反转源句子上训练的 LSTM 在长句子上的表现比在原始源句子上训练的 LSTM 要好得多（见 3.7 节），这表明反转输入句子导致 LSTM 有更好的记忆利用。

3.4 训练细节

我们发现 LSTM 模型相当容易训练。我们使用了 4 层的深层 LSTM，每层有 1000 个单元和 1000 维的词嵌入，输入词汇量为 160,000，输出词汇量为 80,000。我们发现深层 LSTM 在性能上显著优于浅层 LSTM，每增加一层就能将困惑度减少近 10%，可能是因为它们有更大的隐藏状态。我们在每个输出上使用了一个简单的 softmax，覆盖了 80,000 个词。结果的 LSTM 有 3.8 亿参数，其中有 6400 万是纯递归连接（3200 万用于“编码器”LSTM，3200 万用于“解码器”LSTM）。完整的训练细节如下：

- 我们用-0.08 到 0.08 之间的均匀分布初始化了 LSTM 的所有参数。
- 我们使用了没有动量的随机梯度下降，学习率固定为 0.7。5 个周期后，我们开始每半个周期将学习率减半。我们的模型总共训练了 7.5 个周期。
- 我们使用了 128 个序列的批次进行梯度计算，并将其除以批次大小（即 128）。
- 尽管 LSTM 通常不会遭受梯度消失问题，但它们可能会有梯度爆炸。因此，当梯度的范数超过阈值时，我们强制对梯度的范数施加硬性限制[10, 25]。对于每个训练批次，我们计算$s = \|g\|_2$，其中g是除以 128 的梯度。如果$s > 5$，我们设定$g = 5g/s$。
- 不同的句子有不同的长度。大多数句子都很短（例如，长度 20-30），但有些句子很

长（例如，长度>100），因此一个随机选择的包含 128 个训练句子的小批量将有许多短句子和少量长句子，结果是，小批量中的大部分计算是浪费的。为了解决这个问题，我们确保一个小批量内的所有句子长度大致相同，这样可以加速 2 倍。

3.5 并行化

单个 GPU 上使用上一节配置的深度 LSTM 的 C++实现处理速度大约为每秒 1,700 个单词。这对我们的需求来说太慢了，因此我们使用一台 8-GPU 机器对我们的模型进行了并行化。LSTM 的每一层在不同的 GPU 上执行，并且一旦计算出来，就将其激活传递到下一个 GPU（或层）。我们的模型有 4 层 LSTM，每层都驻留在一个单独的 GPU 上。剩余的 4 个 GPU 用于并行化 softmax，因此每个 GPU 负责乘以一个 1000×20000 的矩阵。这种实现的结果是，每秒可处理 6,300 个单词（包括英文和法文），小批量大小为 128。使用这种实现，训练大约需要十天。

3.6 实验结果

我们使用有大小写区分的 BLEU 分数[24]来评估我们翻译的质量。我们使用 multi-bleu.pl[1]在标记化的预测和真实值上计算 BLEU 分数。这种评估 BLEU 分数的方式与[5]和[2]一致，并且再现了[29]的 33.3 分数。然而，如果我们以这种方式评估[9]的最先进系统（其预测可以从 statmt.org\matrix 下载），我们得到的是 37.0，这比 statmt.org\matrix 报告的 35.8 要高。

表 1 LSTM 在 WMT’14 英语到法语测试集（ntst14）上的性能。请注意，使用大小为 2 的 beam 的 5 个 LSTM 的集成比使用大小为 12 的 beam 的单个 LSTM 更节约成本。

方法	BLEU 分数 (ntst14)
Bahdanau 等人. [2]	28.45
基线系统 [29]	33.30
单向正序 LSTM, beam 大小为 12	26.17
单向逆序 LSTM, beam 大小为 12	30.59
5 个逆序 LSTMs 集成, beam 大小为 1	33.00
2 个逆序 LSTMs 集成, beam 大小为 12	33.27
5 个逆序 LSTMs 集成, beam 大小为 2	34.50
5 个逆序 LSTMs 集成, beam 大小为 12	34.81

[1] BLEU 分数有几种变体，每种变体都用 perl 脚本定义。

表 2 在 WMT’14 英语到法语测试集（ntst14）上结合使用神经网络和 SMT 系统的方法。

方法	BLEU 分数 (ntst14)
基线系统 [29]	33.30
Cho 等人 [5]	34.54
State of the art [9]	37.0
Rescoring the baseline 1000-best with a single forward LSTM	35.61
Rescoring the baseline 1000-best with a single reversed LSTM	35.85
Rescoring the baseline 1000-best with an ensemble of 5 reversed LSTMs	36.5
Oracle Rescoring of the Baseline 1000-best lists	~45

结果在表 1 和表 2 中呈现。我们获得的最佳结果是使用一个 LSTM 集成，它们在随机初始化和小批量的随机顺序上有所不同。虽然 LSTM 集成的解码翻译并没有超过最先进的水平，但这是纯神经翻译系统首次在大型机器翻译任务上以可观的差距超过基于短语的 SMT 基线，尽管它无法处理词汇表外的词。通过重新打分基线系统的 1000 个最佳列表，LSTM 的 BLEU 分数接近之前最先进技术 0.5 分。

3.7 长句子上的表现

我们惊讶地发现 LSTM 在长句子上表现良好，这在图 3 中定量显示。表 3 展示了几个长句子及其翻译的例子。

3.8 模型分析

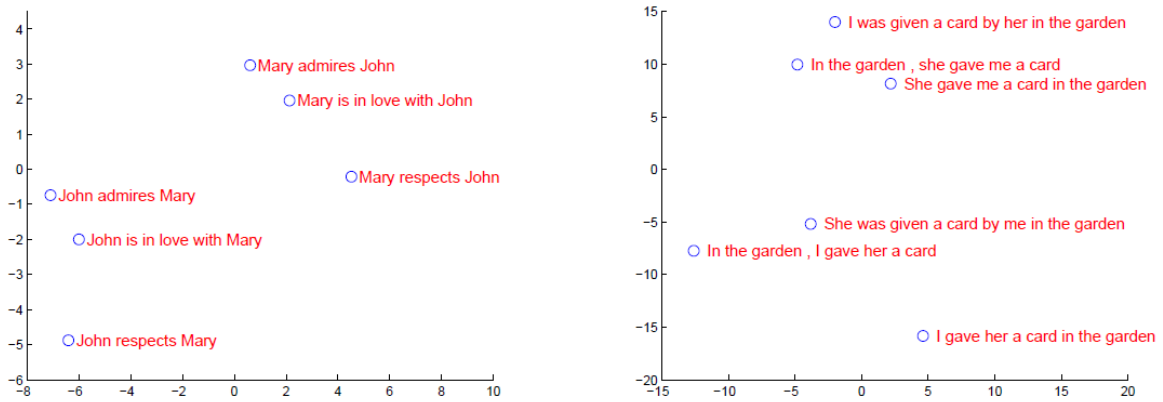

图 2 在处理图中的短语后获得的 LSTM 隐藏状态的二维 PCA 投影。这些短语按照词义进行了聚类，在这些例子中，词义主要是词序的函数，这很难用词袋模型捕捉。两个聚类都有类似的内部结构。

我们模型的一个吸引人的特点是其将一系列单词转换为固定维度向量的能力。

表 3 由 LSTM 生成的一些长句翻译例子以及真实的翻译。读者可以使用谷歌翻译来验证翻译的合理性。

Type	Sentence
Our model	Ulrich UNK , membre du conseil d’administration du constructeur automobile Audi , affirme qu’ il s’ agit d’ une pratique courante depuis des ann´ees pour que les t´el´ephones portables puissent ˆetre collect´es avant les r´eunions du conseil d’ administration afin qu’ ils ne soient pas utilis´es comme appareils d’ ´ecoute `a distance .
Truth	Ulrich Hackenberg , membre du conseil d’administration du constructeur automobile Audi , d´eclare que la collecte des t´el´ephones portables avant les r´eunions du conseil , afin qu’ ils ne puissent pas ˆetre utilis´es comme appareils d’ ´ecoute `a distance , est une pratique courante depuis des ann´ees .
Our model	“ Les t´el´ephones cellulaires , qui sont vraiment une question , non seulement parce qu’ ils pourraient potentiellement causer des interf´erences avec les appareils de navigation , mais nous savons , selon la FCC , qu’ ils pourraient interf´erer avec les tours de t´el´ephone cellulaire lorsqu’ ils sont dans l’ air ” , dit UNK .
Truth	“ Les t´el´ephones portables sont v´eritablement un probleme , non seulement parce qu’ ils pourraient ´eventuellement cr´eer des interf´erences avec les instruments de navigation , mais parce que nous savons , d’ apres la FCC , qu’ ils pourraient perturber les antennes-relais de t´el´ephonie mobile s’ ils sont utilis´es `a bord ” , a d´eclar´e Rosenker .
Our model	Avec la cr´emation , il y a un “ sentiment de violence contre le corps d’ un ˆetre cher ” , qui sera “ r´eduit a une pile de cendres ” en tres peu de temps au lieu d’ un processus de d´ecomposition “ qui accompagnera les ´etapes du deuil ” .
Truth	Il y a , avec la cr´emation , “ une violence faite au corps aim´e ” , qui va ˆetre “ r´eduit a un tas de cendres ” en tres peu de temps , et non apr`es un processus de d´ecomposition , qui “ accompagnerait les phases du deuil ” .

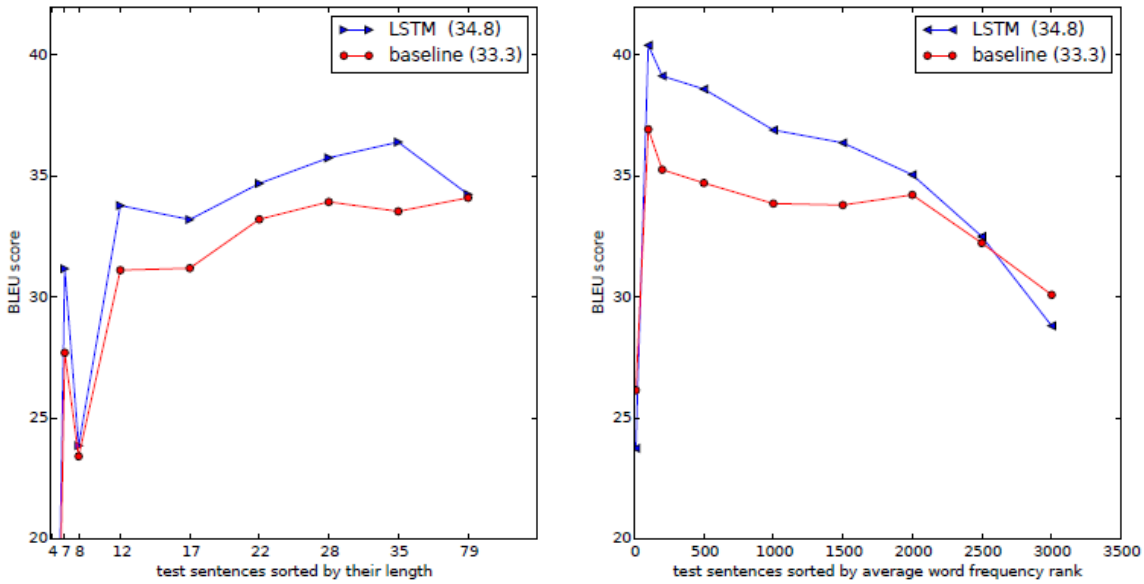

图 2 左图显示了我们系统根据句子长度的性能，其中 x 轴对应按长度排序的测试句子，并标有实际的序列长度。在不超过 35 个单词的句子上没有性能下降，在最长的句子上只有轻微的下降。右图显示了 LSTM 在包含逐渐增加的罕见单词的句子上的表现，其中 x 轴对应按“平均单词频率排名”排序的测试句子。

图 2 可视化了一些学习到的表示。图表清晰显示，这些表示对单词顺序敏感，而对主动语态与被动语态的替换相当不敏感。这些二维投影是使用 PCA 获得的。

4 相关工作

关于将神经网络应用于机器翻译，有大量的研究工作。到目前为止，应用 RNN 语言模型（RNNLM）[23]或前馈神经网络语言模型（NNLM）[3]到机器翻译任务的最简单和最有效的方法是通过对强大的机器翻译基线[22]的 n-best 列表进行重新打分，这可靠地提高了翻译质量。

最近，研究人员开始探索将源语言信息纳入 NNLM 的方法。这方面的工作例子包括 Auli 等人[1]，他们将 NNLM 与输入句子的主题模型相结合，改善了重新打分的表现。Devlin 等人[8]采取了类似的方法，但他们将 NNLM 整合到了机器翻译系统的解码器中，并使用解码器的对齐信息为 NNLM 提供输入句子中最有用的单词。他们的方法非常成功，并在其基线上取得了巨大的改进。

我们的工作与 Kalchbrenner 和 Blunsom[18]密切相关，他们是首次将输入句子映射成向量然后再映射回句子的，尽管他们使用卷积神经网络将句子映射到向量，这丢失了单词的顺序。与此工作类似，Cho 等人[5]使用了一个类似 LSTM 的 RNN 架构将句子映射成向量并再映射回来，尽管他们的主要关注点是将他们的神经网络整合到一个 SMT 系统中。Bahdanau 等人[2]也尝试了用一个使用注意力机制的神经网络进行直接翻译，以克服 Cho 等人[5]在长句子上的表现不佳，取得了鼓舞人心的结果。同样，Pouget-Abadie

等人[26]试图解决 Cho 等人[5]提到的记忆问题，通过翻译源句子的片段以产生流畅的翻译，这与基于短语的方法类似。我们怀疑他们通过简单地在反转的源句子上训练他们的网络可以取得类似的改进。

端到端训练也是 Hermann 等人[12]的研究重点，他们的模型通过前馈网络表示输入和输出，并将它们映射到空间中的相似点。然而，他们的方法不能直接生成翻译：要得到一个翻译，他们需要在预先计算好的句子数据库中查找最近的向量，或重新打分一个句子。

5 结论

在这项工作中，我们展示了一个大型的深层 LSTM，尽管其词汇量有限，但可以在一个大规模的机器翻译任务上超过词汇量无限的标准基于 SMT 的系统。我们基于简单的 LSTM 方法在机器翻译上的成功表明，只要有足够的训练数据，它应该在许多其他序列学习问题上也能做得很好。

我们对反转源句子中的单词所获得的改进程度感到惊讶。我们得出结论，寻找具有最多短期依赖关系的问题编码是很重要的，因为它们使学习问题大为简化。特别是，虽然我们无法训练一个标准 RNN 来处理未反转的翻译问题（如图 1 所示），我们相信当源句子被反转时，应该很容易训练一个标准 RNN（尽管我们没有在实验中验证它）。

我们也对 LSTM 正确翻译非常长的句子的能力感到惊讶。我们最初确信由于其有限的记忆，LSTM 会在长句子上失败，其他研究人员也报告了与我们的模型类似的在长句子上的表现不佳[5, 2, 26]。然而，经过反转数据集训练的 LSTMs 在翻译长句子时几乎没有困难。

最重要的是，我们证明了一个简单、直接和相对未优化的方法可以超过成熟的 SMT 系统，因此后续工作很可能会带来更高的翻译准确性。这些结果表明，我们的方法可能会在其他具有挑战性的序列到序列问题上表现良好。

6 致谢

我们感谢 Samy Bengio, Jeff Dean, Matthieu Devin, Geoffrey Hinton, Nal Kalchbrenner, Thang Luong, Wolfgang Macherey, Rajat Monga, Vincent Vanhoucke, Peng Xu, Wojciech Zaremba, 以及 Google Brain 团队对有用的评论和讨论。

参考文献

[1] M. Auli, M. Galley, C. Quirk, and G. Zweig. Joint language and translation modeling with recurrent neural networks. In EMNLP, 2013.

[2] D. Bahdanau, K. Cho, and Y. Bengio. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473, 2014.

[3] Y. Bengio, R. Ducharme, P. Vincent, and C. Jauvin. A neural probabilistic language model. In Journal of Machine Learning Research, pages 1137－1155, 2003.

[4] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. IEEE Transactions on Neural Networks, 5(2):157－166, 1994.

[5] K. Cho, B.Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Arxiv preprint arXiv:1406.1078,2014.

[6] D. Ciresan, U. Meier, and J. Schmidhuber. Multi-column deep neural networks for image classification. In CVPR, 2012.

[7] G. E. Dahl, D. Yu, L. Deng, and A. Acero. Context-dependent pre-trained deep neural networks for large vocabulary speech recognition. IEEE Transactions on Audio, Speech, and Language Processing - SpecialIssue on Deep Learning for Speech and Language Processing, 2012.

[8] J. Devlin, R. Zbib, Z. Huang, T. Lamar, R. Schwartz, and J. Makhoul. Fast and robust neural network joint models for statistical machine translation. In ACL, 2014.

[9] Nadir Durrani, Barry Haddow, Philipp Koehn, and Kenneth Heafield. Edinburgh's phrase-based machine translation systems for wmt-14. In WMT, 2014.

[10]A. Graves. Generating sequences with recurrent neural networks. In Arxiv preprint arXiv:1308.0850,2013.

[11]A. Graves, S. Fern´andez, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In ICML, 2006.

[12]K. M. Hermann and P. Blunsom. Multilingual distributed representations without word alignment. In ICLR, 2014.

[13]G. Hinton, L. Deng, D. Yu, G. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen,T. Sainath, and B. Kingsbury. Deep neural networks for acoustic modeling in speech recognition. IEEE Signal Processing Magazine, 2012.

[14]S. Hochreiter. Untersuchungen zu dynamischen neuronalen netzen. Master's thesis, Institut fur Infor-matik, Technische Universitat, Munchen, 1991.

[15]S. Hochreiter, Y. Bengio, P. Frasconi, and J. Schmidhuber. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies, 2001.

[16]S. Hochreiter and J. Schmidhuber. Long short-term memory. Neural Computation, 1997.

[17]S. Hochreiter and J. Schmidhuber. LSTM can solve hard long time lag problems. 1997.

[18]N. Kalchbrenner and P. Blunsom. Recurrent continuous translation models. In EMNLP, 2013.

[19]A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet classification with deep convolutional neural networks. In NIPS, 2012.

[20]Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, and A.Y. Ng. Building high-level features using large scale unsupervised learning. In ICML, 2012.

[21]Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998.

[22]T. Mikolov. Statistical Language Models based on Neural Networks. PhD thesis, Brno University of Technology, 2012.

[23]T. Mikolov, M. Karafi´at, L. Burget, J. Cernock`y, and S. Khudanpur. Recurrent neural network based language model. In INTERSPEECH, pages 1045–1048, 2010.

[24]K. Papineni, S. Roukos, T. Ward, and W. J. Zhu. BLEU: a method for automatic evaluation of machine translation. In ACL, 2002.

[25]R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. arXiv preprint arXiv:1211.5063, 2012.

[26]J. Pouget-Abadie, D. Bahdanau, B. van Merrienboer, K. Cho, and Y. Bengio. Overcoming the curse of sentence length for neural machine translation using automatic segmentation. arXiv preprint arXiv:1409.1257, 2014.

[27]A. Razborov. On small depth threshold circuits. In Proc. 3rd Scandinavian Workshop on Algorithm Theory, 1992.

[28]D. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. Nature, 323(6088):533–536, 1986.

[29]H. Schwenk. University le mans. http://www-lium.univ-lemans.fr/~schwenk/cslm_joint_paper/, 2014. [Online; accessed 03-September-2014].

[30]M. Sundermeyer, R. Schluter, and H. Ney. LSTM neural networks for language modeling. In INTER-SPEECH, 2010.

[31]P. Werbos. Backpropagation through time: what it does and how to do it. Proceedings of IEEE, 1990.